Semantic User Interaction Profiles for Better People Recommendation

Johann Stan*†, Viet-Hung Do*
*Alcatel-Lucent Bell Labs France
Centre de Villarceaux
firstname.lastname@alcatel-lucent.com

Pierre Maret†
†Université de Lyon, F-42023
CNRS, UMR 5516, Laboratoire Hubert Curien,
Université de Saint-Etienne, Jean-Monnet
firstname.lastname@univ-st-etienne.fr

***Abstract*—In this paper we present a methodology for learning user profiles from content shared by people on Social Platforms. Such profiles are specifically tailored to reflect the user's degree of interactivity related to the topics they are writing about. The main novelty in our work is the introduction of Linked Data in the content extraction process and the definition of a specific scores to measure expertise and interactivity.**

I. Introduction

The analysis of shared content on social platforms may provide a new input for advanced recommendation strategies, as it offers valuable insights into people's interests, plans, findings and information needs. In recent years there has been a lot of investement to understand the structure of such systems, especially social networks. [**?**] points out that although important results have been achieved (small worlds, community extraction, key player identification), the analysis of shared content could be a significant added-value for advanced recommendation strategies built on top of such systems.

In this paper we present a framework for building *User Interaction Profiles (UIP)* from content shared by users (tags, microsposts) on social platforms and we describe a prototype implementation for advanced people recommendation during web navigation. A first specificity of UIPs is that they are constructed from content productions instead of consumptions, which allows better respect of privacy: we only process information already explicitly shared by users. Also, content productions are more representative of what people's activities, likes and dislikes. The second relates to the different scores we associate to each profile concept, that attempt to measure the degree of expertise and interactivity (e.g. high sentiment expressed in a message can indicate higher motivation to discuss the underlying topics). In the following section we study the content sharing habits of users on social platforms with a specific focus on Twitter. In section 2 we introduce our method for the extraction of meaningful concepts from such messages and the formal definition of interaction profiles. In section 3 we consider the issue of weighting concepts in the interaction profile in order to measure the user's interactivity. This is followed by the description of a working prototype. Finally, in section 4 we present related work followed by a conclusion and future plans in section 5.

Fig. 1. Semantic structure of social updates

II. The User Interaction Profile Builder Framework

In recent years, social platforms (e.g. social networking sites, microblogging sites, bookmarking sites) have democratized the way people share information about their activity, findings or information needs. A new form of content sharing has emerged, replacing large, authoritative documents with short, unstructured, amateur free-form text. This spans a wide range from tags used for photo annotations to microposts that translate the success of SMS to content sharing on the Web. These are mainly free text annotations that describe resources, activities or feelings [**?**]. We analyzed 45000 social updates shared by 830 users on Twitter to better understand the composition of such messages by focusing on the presence of named entities (either directly or indirectly in the form of URLs that point to web pages). In Figure 1 we represented the total number of shared messages and those containing named entities for each user.

As it can be observed, three categories of users can be distinguished. A first category is composed of users who share messages containing only keywords, probably for communicating their activity in a very general way (such as "I like painting"). The second category includes users who share more precise and specific information dominated by the presence of named entities or URLs (such as "I like Claude Monet"). Finally, there is a third category doing both. The strategy to capture user interests from shared content must therefore be adapted to all categories of users. We leverage Linked Data to tackle this issue. Linked Data is a community effort to extract data from various sources, interlink it and store it in a semantic web format, like RDF. Currently, the most complete dataset is DBPedia[**?**], with more than 3.5 million concepts and relations extracted from Wikipedia. Such knowledge bases represent a valuable resource that allows both generalization and specialization of concepts. The underlying data structure is represented by a directed graph, where nodes are composed and edges properties that connect them (e.g. Paris isCapitalOf France).

Fig. 2. Semantic Framework for Social Search

A. General Description of the Framework

The framework (Figure 2) is composed of three layers: (i) Aggregator, (ii) Analysis and (iii) Semantic Service layers.

- *Aggregator Layer.* Its main role is the crawling of shared social updates of a given user in the social platform. This component implements the API offered by each social platform (1). This mainly consists of the implementation of the authentification mechanism and the extraction of interactions on a regular basis, depending on the limitations of the API. Interactions are stored in a relational database (Aggregation - A).
- *Analysis Layer.* The analysis of the retrieved social updates. The analysis includes methods for keyword and entity extraction, disambiguation and concept expansion. These operations compose the User Model Builder Engine (2). This layer also includes the different interactivity measures (Scoring Engine - 3), user feedback and ranking engines (5). A specific component of the Analysis Layer is the Linked Data store, in our case a dump of DBPedia (4). Its main role is to provide a structured semantic layer to the extracted content. The disambiguated structured data is reinjected in a Sesame data store[1] (B).
- *Service Layer.* The construction of the user interaction profiles and corresponding services that explore them. Each service queries the Sesame data store and extracts the information that it needs (C).

B. From Social Updates to Linked Data Concepts

The process of linking keywords and named entities extracted from social updates to Linked Data concepts is composed of the following steps: (i) keyword, entity extraction, (ii) disambiguation and (iii) semantic expansion. For the first operation we use an existing service, AlchemAPI[2], which employs statistical algorithms to find the most relevant terms in a text (tf-idf, part-of-speech tagging, word-cooccurences).

1) Disambiguation of Social Updates: The main difficulty in connecting named entities and keywords extracted from posts shared by the user to Linked Data concepts is the choice of the best concept from the knowledge base that best approximates their meaning. In order to associate keywords or entities in a social update to the right concept in Linked Data, contextual cues are necessary to allow restricting the semantic field of the social update (e.g. the social update "I like apple" can both refer to apple as a fruit and a company). In traditional documents, generally there are sufficient contextual cues to overcome such ambiguous situations, where the meaning of the message is not straightforward. In our case, the short nature of social updates requires us to find such cues elsewhere. In our framework we consider two main additional sources of contextual cues: user-related, which consists in building incrementally a vocabulary from all social updates of the user. The assumption behind this first additional context is that there is a probability that the user previously shared some content in a related semantic field (e.g. a user who posted about the iPhone might have shared before about other Mac products). The second additional contextual cue comes from the community of the user. On social platforms users are members of different communities, which influence each other in terms of interests (phenomenon called homophily). Once this contextual vocabulary is constructed, for each message posted by the user, a set of candidate concepts are retrieved, based on the similarity of its label and the entity to be disambiguated in the message. Secondly, the abstract of each candidate concept is compared to the contextual cue vocabulary with cosine similarity and the most similar considered as the winner concept.

2) Semantic Expansion using Linked Data: Once this first association has been performed, we have to face the second difficulty: the semantic expansion of the concept. The main role of this operation is to allow us better categorizing the user's profiles and to better approximate interests. As an example, a user who shared about topics such as Facebook and Twitter might have a general interest in Web 2.0 technologies, which can be inferred by propagating the atomic interests to the more general categories. The first step in this operation consists in building a semantic sphere associated to the initial concept (Figure 3), that contains all candidate concepts that will form the expansion set. In our approach, we explore three types of connections in Linked Data to construct this shpere: the first and the most interesting is represented by hierarchical links to category concepts (e.g. concept "Gran Torino" will have "Gang films" and "American drama films" as hierarchical expansions). They generalize a concept to categories, e.g. someone interested in the film Gran Torino may also be interested in "Gang films". In the second case, we explore concepts connected to each category concept that was previously retrieved (e.g. the film Punisher, the neighbor of Gran Torino in the Gang films category). The third dimension explores concepts directly connected to the initial concept in the knowledge base (e.g. Clint Eastwood is the director of Gran Torino). According to the structure of shared content (general or specific, as seen in Section 2), more weight can be given to specific dimensions (e.g. if the user shares only general messages, more concepts will be explored in the category and direct dimensions).

The set of properties P which is chosen from the knowledge base properties, allows the system to expand a concept in three predefined dimensions. The property "*subject*" of a concept c returns its category which corresponds to the hierarchy dimension. The property "*property*" returns the concept from its infobox, corresponding to the direct dimension. Finally, the property "*isbroaderof(subject(c))*" corresponds to the concepts that are in the same category as the initial concept.

3) Concept Filtering: The final step of this expansion is the filtering. A knowledge base often contains a very large amount

[1]Sesame Datastore - www.openrdf.org- visited January 2011
[2]Alchemy API - www.alchemyapi.com/ - visited March 2011

Fig. 3. Dimensions of the semantic expansion of concept "Gran Torino" - a movie directed by Clint Eastwood in 2010

Fig. 4. User Interaction Profile Manager

of data from the very specific to the very general. Therefore the result set obtained from a semantic expansion operation is always large (represented by the external sphere in 3. In our approach we want to keep only a subset of concepts that are the closest ones to the user's interests (inner sphere in Figure 3). To perform the filtering, we compute the similarity of the name of each concepts present in the expansion with the abstracts of the concepts associated to the user's social update. An abstract of a concept is a small paragraph available for every concept in a Linked Data knowledge base that gives its definition. The fact that we compute the similarity score with the abstracts is based on the idea that an abstract will normally contain lots of keywords related to the concept. These keywords build up a small vocabulary which can serve like a local context in order to find the closest concepts from the expansion (e.g. The abstract of the film Gran Torino contains essentially keywords like "American", "drama" and more frequently "Eastwood"). When compared to this local vocabulary, the concepts like "American drama" or "Clint Eastwood" will certainly score more points in similarity than the others like "Fictional American people of Polish descent", also present in the expansion set. Once we have the similarity scores, we rank the concepts by sorting them and add to user's profile the concepts with the top-k highest scores (k is set to 5 in the current implementation).

C. Interactivity Measures

In order to understand how users express interactivity in their social updates, we followed an experimentation protocol with a set of 20 users (students and researchers). Users were asked to log in the system with their Twitter account[3]. The system processed the 60 last social updates of each user and extracted named entities and keywords. Using the concept tagging, we generalized each of them with category concepts from DBPedia in order to show users a more structured profile that is easier to navigate and understand. The profile was shown in the form of a tag cloud (Figure 4 - User Interaction Profile at the left). Each user was asked to rate concepts in the tag cloud with marks ranging from 1 to 10 according to their degree of interactivity associated (e.g. to rate their motivation to answer questions or enagage in conversation with that specific topic) (Figure 4). We examined the social updates that contain concepts that were given higher scores than 5. The main observations were the following: (i) social updates relevant to these concepts express generally high sentiment (negative or positive) and (ii) messages in this category were of better quality than other updates from the same user. Therefore, we implemented a ranking function that considers the sentiment polarity and the entropy of social updates related to a given concept. We compute the sentiment polarity of interactions by first performing the part-of-speech tagging of messages and then retrieving the polarity of the keywords with the SentiWordNet [**?**] vocabulary. The matching of the keyword to the right synset is performed by taking into account the grammatical correspondence. This approach seemed more efficient than other machine-learning based algorithms, given the short nature of message and the relatively week vocabulary used by users to express sentiments. In case of the second criteria, we use entropy to measure the richness of messages related to a given topic.

[3]Bell Labs Social Search Engine - www.codex-project.com/semantic

Fig. 5. Social Adviser: Connecting social platforms to web navigation.

III. Example Prototype: the Social Adviser

The Social Adviser (Figure 5) is a prototype application implemented as an extension of the web browser that consumes the Service layer of the framework (Rest interface). The main idea is to consider the currently viewed web page of the user as a query to the social network. This is achieved by the extraction of entities and keywords from the web page, their disambiguation using DBPedia and the construction of a first semantic tag cloud from the data (i.e. the query). This expanded tag cloud is further compared to the interaction profiles of users in the social network and the top 5 most similar users are shown. Thus, users who in their interactions relevant to the query expressed strong sentiment (positive or negative) will be shown first. The user has then several options: (i) check the interactions of the recommended user or (ii) contact her. Also, a summarized view of the recommended users is shown in the form of a graph, with distances based on the semantic similarity to the query concept and colors representing the associated interactivity (colors in case of sentiment (green-positive, red-negative or grey in case of high entropy). In the example scenario shown in Figure 5, the user is currently viewing a page about the movie "Gran Torino". The Social Advizer shows friends who posted relevant social updates with high interactivity about this topic (e.g. high sentiment or high entropy). Also, if a new social update relevant to this item is posted by friends, the Social Adviser automatically updates the output.

IV. Related Work

Our work is related to the area of resource discovery in the case of social content sites [**?**]. The corresponding operation is called social search and consists in retrieving social resources instead of web documents. A particular case of social search that is closest to our work consists in people recommendation. *Aardvark* [**?**] is a social system allowing users to define a profile and ask questions in natural language and sends it to users with related topics in their profile. In our framework we

propose a different approach, by exploring directly the social network of the user for recommendations and using Linked Data to better approximate user interests. Also, we use specific scores to find users who shared interesting messages, either with high sentiment or entropy. The content of social updates for recommendation was also explored in the Twittomender system [**?**]. Designed for recommending people to follow in Twitter, the *Twittomender* system allows users to expand their network by connecting to people that they don't know directly, but with whom they share similar interests. Each user in the system is represented by a bag-of-words, comprised of terms extracted from their shared messages. Another recent work explored the issue of people recommendation but considering only the the frequency of interactions (phone calls) [**?**].

V. Conclusion and Future Work

In this paper we presented a framework that allows transforming a social platform into a social search engine. At the basis of this transformation are so called User Interaction Profiles, constructed from the shared content of the user. The main innovation of our work is the combination of traditional content analysis techniques (entity, keyword extraction) with semantic web technologies (Linked Data). A first long-term objective of this work is to reduce to knowledge latency in social platforms, which we define as the time needed to find the most relevant and useful knowledge for a given information need.

References